Um sistema de recomendação para conteúdos da cultura da cana-de-açúcar

Flávio Margarito Martins de Barros[1]
Stanley Robson de Medeiros Oliveira[2]
Leandro Henrique Mendonça de Oliveira[2]

Sistemas de informação web oferecem informações em quantidade elevada, tanto que a tarefa de encontrar os dados de interesse torna-se desafiadora. A Agência de Informação Embrapa é um sistema web que tem como objetivo: organização, tratamento, armazenamento e divulgação de informações técnicas e conhecimentos gerados pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa). O portal está estruturado como uma árvore hierárquica, denominada Árvore de Conhecimento, a qual compreende centenas de páginas web, artigos, planilhas e materiais multimídia. Diariamente, o site recebe milhares de acessos que são registrados em uma base de dados. Nesse domínio, onde temos informações em quantidade elevada, armazenadas digitalmente em bases de dados, as ferramentas de Mineração de Dados apresentam recursos para análise que podem fornecer padrões de uso do site para fazer recomendações. Recomendações personalizadas de conteúdo aumentam a usabilidade dos sistemas, agregam valor aos serviços, poupam tempo e fidelizam usuários. O objetivo desse trabalho foi projetar e desenvolver um sistema de recomendação web, baseado em regras de associação, que ofereça recomendações automaticamente de conteúdos da cultura da cana-de-açúcar, de acordo com o perfil dos usuários.

A metodologia utilizada na pesquisa seguiu o modelo CRISP-DM (CHAPMAN et al., 2000), composta por seis etapas, a saber: a) compreensão do domínio (perfis de acesso sobre páginas de cana-de-açúcar; b)

[1] Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola, flavio.barros@feagri.unicamp.br

[2] Embrapa Informática Agropecuária, {stanley.oliveira, leandro.oliveira}@embrapa.br

entendimento dos dados (registros de tráfego de internet armazenados em duas tabelas: clientes com 919.212 linhas e a tabela *tracker* com 1.990.616 linhas, com informações dos usuários e dos acessos, respectivamente.); c) preparação dos dados (transformação de registros de acesso para uma estrutura de "lista de acessos", onde estão registradas todas as páginas visitadas pelo usuário); d) modelagem (geração de regras de recomendação por meio do algoritmo Apriori); e) avaliação (regras mais importantes são retidas); f) distribuição (ocorre por meio de links oferecidos como recomendações de leitura).

A partir dos dados armazenados no banco de dados, regras de associação entre páginas foram geradas com o algoritmo *Apriori* (AGRAWAL et al., 1993). Regras de associação descrevem a relação entre itens ou produtos de uma base de dados. Essas regras podem ser representadas da forma $X \rightarrow Y$, onde X e Y são conjuntos disjuntos de atributos, isto é, $X \cap Y = \emptyset$. Nessas regras, X representa o antecedente e Y, o consequente. Para o *ranking* das regras de associação foi utilizada uma métrica denominada MaxConf (HAN et al., 2011). Essas regras foram geradas utilizando a linguagem R e o pacote *arules* (HAHSLER; GRUEN, 2011) e armazenadas em um sistema gerenciador de banco de dados[3], PostgreSQL 8.4 para armazenamento e consulta dos dados.

A estrutura do sistema, apresentada na Figura 1, compreende o equipamento onde estão instalados o servidor de páginas, o banco de dados, o

Figura 1. Arquitetura do sistema de recomendação.

[3] Disponível em: <http://www.postgresql.org.br/>.

sistema de recomendação nra forma de um *script* em R e a comunidade de usuários. Os acessos dos usuários alimentam o banco de dados, de onde o sistema de recomendação extrai os padrões de acesso da comunidade. Sempre que um usuário interage com o sistema essa interação é armazenada no banco e o sistema é retroalimentado com as recomendações sendo melhoradas no processo.

Se um usuário, por exemplo, acessa a página “Cana-de-açúcar”, o sistema possui quatro possíveis recomendações (as quatro primeiras linhas da Tabela 1). Assim o usuário tem acesso às páginas mais fortemente associadas em relação aos padrões de uso da comunidade de visitantes do portal.

Tabela 1. Regras geradas pelo algoritmo *Apriori* para recomendação de páginas.

Antecedente	Consequente	MaxConf
Cana-de-açúcar	Causas de acidentes	96,40%
Cana-de-açúcar	Picagem da cana	95,93%
Cana-de-açúcar	Recomendações gerais	95,65%
Cana-de-açúcar	Resultados alcançados	94,91%
Manejo do solo	Amostragem de solo	94,21%

As regras que compõem a base de conhecimento são o resultado da extração dos padrões de uso de muitos usuários da comunidade. Esses padrões refletem importantes associações entre páginas que podem não estar explicitadas na estrutura de links do portal. Assim usuários mais avançados, ao acessarem e gerarem esses padrões auxiliam usuários menos experientes a encontrar informações relevantes utilizando essas recomendações. O objetivo do sistema de recomendação proposto foi transferir conhecimento sobre o uso do portal para a comunidade. Esse conhecimento foi sumarizado e armazenado em um banco de dados na forma de regras de associação entre páginas.

Referências

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. N. Mining Association Rules between Sets of Items in Large Databases. **SIGMOD**, Washington, v. 22, n. 2, p. 207-216, 1993.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0**: step-by-step data mining guide. Illinois: SPSS, 2000. 78 p.

HAHSLER, M.; GRUEN, B.; HORNIK, K. **Arules**: Mining Association Rules and Frequent Itemsets. R package version 1.0-8 , 2011.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining**: Concepts and Techniques, 3rd. Burlington: Elsevier, 2011. 703 p.